Structure, Vol. 13, 121–130, January, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.str.2004.10.015

Inference of Protein Function from Protein Structure

Debnath Pal and David Eisenberg*
UCLA-DOE Institute for Genomics and Proteomics
Howard Hughes Medical Institute
Box 951570
Los Angeles, California 90095

Summary

Structural genomics has brought us three-dimensional structures of proteins with unknown functions. To shed light on such structures, we have developed ProKnow (http://www.doe-mbi.ucla.edu/Services/ProKnow/), which annotates proteins with Gene Ontology functional terms. The method extracts features from the protein such as 3D fold, sequence, motif, and functional linkages and relates them to function via the ProKnow knowledgebase of features, which links features to annotated functions via annotation profiles. Bayes' theorem is used to compute weights of the functions assigned, using likelihoods based on the extracted features. The description level of the assigned function is quantified by the ontology depth (from 1 = general to 9 = specific). Jackknife tests show 89% correct assignments at ontology depth 1 and 40% at depth 9, with 93% coverage of 1507 distinct folded proteins. Overall, about 70% of the assignments were inferred correctly. This level of performance suggests that ProKnow is a useful resource in functional assessments of novel proteins.

Introduction

A major goal of molecular biology is to understand functions of all genes in nature. Structural genomics initiatives contribute significantly toward this goal by producing three-dimensional structures of many proteins, which allow us to better understand sequence-structure-function relationships. But knowing sequence and structure does not guarantee knowing protein function, especially in cases where there is no history of experimental characterization. Over time, large-scale functional genomics/proteomics experiments will fill the gaps. Meanwhile, in silico methods capable of function annotation of proteins must be extended.

The word "function" within a biological context is an evolving concept and is used in many ways. Webster's Dictionary describes function as "any of a group of related actions contributing to a larger action, especially: the normal and specific contribution of a bodily part to the economy of a living organism." This definition implies that although functions occur at many levels in an organism (such as molecule, organelle, cell, tissue, organ, and organism), none of them is in isolation. Lower-level functions work together to produce a higher-level function. Also, a lower-level function can be part of many different higher-level functions. The interactions between these functions form the basis for sustainable homeostasis. These multiple levels of function are reflected in our procedure, described below, of linking protein features to annotations at various levels.

The repertoire of methods for in silico annotation of function has grown enormously over the past two decades. A protein with a high degree of sequence similarity to a family of well-characterized proteins can be detected by BLAST (Altschul et al., 1990). With lower sequence similarity, more subtle methods such as "profiles" (where patterns obvious from multiple sequence alignment are evident) (Altschul et al., 1997; Bork and Gibson, 1996; Gribskov et al., 1987) or hidden Markov models (HMM) (Eddy et al., 1995) are required. These methods are based on the assumption that similar sequences have descended from a common ancestor and share similar function. The assumption is, however, limited in validity, as demonstrated by numerous studies (Devos and Valencia, 2000; Gerlt and Babbitt, 2000; Karp, 1998; Rost, 2002; Rost et al., 2003; Rost and Valencia, 1996; Tian and Skolnick, 2003; Whisstock and Lesk, 2003). To enhance accuracy of functional assignment, functional annotations can be inferred from information on fold (Bowie et al., 1991; Holm and Sander, 1998; Jones et al., 1992), motif (Attwood et al., 2003; Henikoff et al., 2000; Hulo et al., 2004), domain (Bateman et al., 2004), and orthology (Tatusov et al., 1997). Another class of annotation algorithms infers protein function based on identification of functionally significant residues. This class includes biodictionary "seqlets" mapping sequence patterns to their properties (Rigoutsos et al., 2002), evolutionary tracing (Landgraf et al., 2001; Yao et al., 2003), graph theory (Wangikar et al., 2003), clique detection (Schmitt et al., 2002), and 3D template matching (Wallace et al., 1996). In all instances, some prior knowledge of sequence or structural similarity is essential for any inference. Support vector machines based on residue properties such as hydrophobicity, polarity, polarizability, solvent accessibility (Cai et al., 2003), or neural networks trained on protein features (Jensen et al., 2003) are some recent approaches to detect function, adding information to basic sequence and structure. The success of these methods, though encouraging, is limited in coverage and accuracy.

Recent advances in our understanding of proteins have revealed new facets of protein function. Moonlighting proteins have been discovered whose functions depend on cellular context (Jeffery, 1999). Even proteins with the same fold and active site architecture have been found with different functions (Wise et al., 2002). Another recent development is the attempt to understand protein function by placing a protein in its cellular context (Eisenberg et al., 2000). These new facets need to be addressed in inferring protein function.

Here, we present a metaserver named ProKnow, which annotates function based on features of protein such as its 3D fold, sequence, structural and sequence motifs, and functional linkages. The backbone of ProKnow is the ProKnow knowledgebase of protein fea-

*Correspondence: david@mbi.ucla.edu

Table 1. Subdatabases of the ProKnow Knowledgebase

	File Name	Description	Number
Downloaded files (from http://www.geneontology.org, http://www.expasy.ch)	SWISS-PROT.GOA	GO annotations for SWISS-PROT	3,032,146 annotations
	SWISS-PROT_FASTA	FASTA format protein sequence from SWISS-PROT and TREMBL	129,463 sequences
	TREMBL_FASTA		855,779 sequences
	TREMBL_NEW_FASTA		190,164 sequences
Knowledgebase A (IEA+, electronic annotations included)	GOSPTR-A	GO annotations for sequence	655,244 sequences
	GOPDB-A	GO annotations for PDB based on fold	30,345 protein chains
	GOPROSITE-A	GO annotations for sequence motifs	949,090 motifs
	GORIGOR-A	GO annotations for 3-dimensional motifs	10,230 3D motifs
	GODIP-A	GO annotations for DIP proteins	3,146 proteins
Knowledgebase B (IEA-, electronic annotations excluded)	GOSPTR-B	GO annotations for sequence	16,441 sequences
	GOPDB-B	GO annotations for PDB based on fold	7,887 protein chains
	GOPROSITE-B	GO annotations for sequence motifs	136,861 motifs
	GORIGOR-B	GO annotations for 3D motifs	7,619 3D motifs
	GODIP-B	GO annotations for DIP proteins	1,973 proteins

Files in the ProKnow knowledgebase were derived from the SWISS-PROT.GOA file. For example, in the SWISS-PROT_FASTA file, which was used to compile PSI-BLAST query database, only those sequences which had annotation in SWISS-PROT.GOA were taken. Similarly, all motifs culled from sequences present in SWISS-PROT.GOA were used to construct the GOPROSITE database. Knowledgebase A is normally used by ProKnow; knowledgebase B was used during evaluation on test set B.

tures. In this knowledgebase, each protein feature is associated with all potential functions (Table 1). We call the collection of functions associated with a protein feature an annotation profile (Supplemental Table S1). When a protein is submitted to ProKnow (Figure 1), the server extracts all identifiable features of the protein. ProKnow then looks to its knowledgebase to map matching protein features, which give the annotation profiles for the query protein. The functions in the mapped profiles that are linked to most protein features are then culled and weighted by Bayes' theorem (Pitman, 1997) for functional assignments using Gene Ontology (GO) terms (Gene Ontology Consortium, 2001). The GO terms are unique numeric labels that represent controlled vocabularies arranged as ontologies that describe function in a hierarchy of directed acyclic graphs

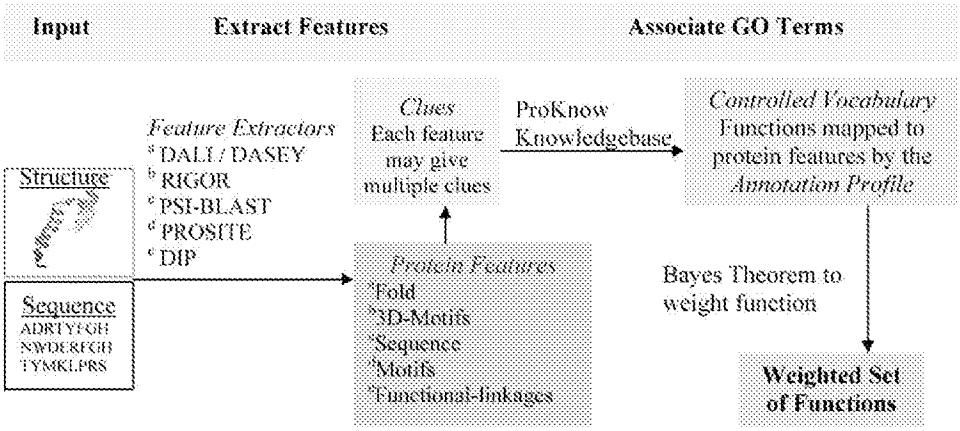


Figure 1. Flowchart for ProKnow

The method can take either a single sequence or a 3D structure as input. The scheme has three major steps, shown in the top panel. Individual steps under these are shown in separate shaded boxes. Protein features and the algorithms that extract them are given the same letter (a to e) (DALI [Holm and Sander, 1998], DASEY [Mallick et al., 2002], RIGOR [Kleywegt, 1999], PROSITE [Hulo et al., 2004], PSI-BLAST [Altschul et al., 1997], DIP [Xenarios et al., 2002]). For input of a protein structure, all protein features are queried, whereas features labeled a, c, d, and e are queried for protein sequence alone. Functional linkages are obtained from the DIP database through proteins linked by a single edge to the proteins obtained by PSI-BLAST search. All feature-extracting programs are used at their default parameter values. More than one clue for a function can be obtained from a given protein feature.

Table 2. GO Evidence Codes and Their Assigned Ranks

Evidence Description	GO Evidence Code	Rank
Inferred by curator	IC	0
Traceable author statement	TAS	1
Inferred from direct assay	IDA	1
Inferred from mutant phenotype	IMP	2
Inferred from genetic interaction	IGI	2
Inferred from physical interaction	IPI	2
Inferred from sequence or structural similarity	ISS	3
Inferred from expression pattern	IEP	3
Nontraceable author statement	NAS	4
Inferred from electronic annotation	IEA	5
No data	ND	6
No record	NR	6

Ranks indicated in this table are used as numeric counterpart to the alphabetic evidence codes supplied with each annotation by the GO consortium. The ranks are empirically assigned by the authors based on intuitive measure of reliability. Evidence rank (ER) used in the text is calculated from the rank values shown in this table. ER is a measure of the quality of the assigned function terms based on the averaged rank of the evidence code of the GO terms used for making the assignments; ER ranges from 0 (best) to 6 (worst). ER is calculated as: ER = (sum of the ranks of N GO terms used in function assignment)/N.

(DAG) (explained in Supplemental Figure S1A). The GO function can be of two types, molecular function or a biological process. A "molecular function" is defined as what a protein does at the biochemical level, while "biological process" refers to a biological objective to which a protein contributes. The description level of the assigned GO function is quantified by the ontology depth (from 1 = general to 9 = specific). Jackknife tests on ProKnow show about 85% correct assignments at ontology depth 1 and 40% at depth 9, with 93% coverage of the molecular function annotations for 1507 distinct folded proteins. Overall, about 70% of the assignments were inferred correctly. Below, we describe the use and performance of ProKnow, available at http://www.doe-mbi.ucla.edu/Services/ProKnow/, to assess GO functions of novel proteins.

Results

The output of ProKnow consists of GO terms, each with its associated Bayesian weight (BW), evidence rank (ER), and clue count (CC). BW indicates the probability of the function (represented by GO term) based on the protein features; BW ranges from 0 to 1. ER is a measure of the quality of the assigned GO terms based on the averaged rank of the evidence code of GO terms used for the GO assignments; ER ranges from 0 (best) to 6 (worst) (Table 2). CC is the number of weights derived from the protein features that were used to calculate BW; the values range from 1–9. A full CC set contains two weights, each computed from 3D fold, sequence, sequence motif, 3D motif, and one from functional linkage.

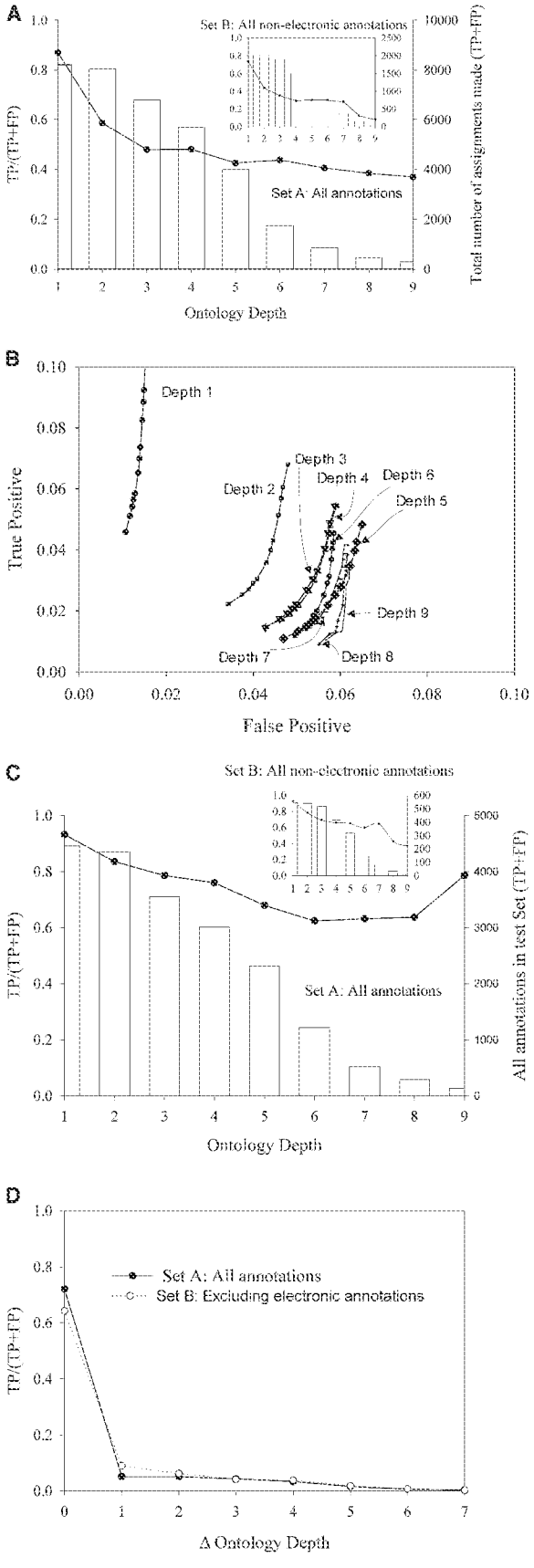
To evaluate the results from ProKnow, we took two sets of Protein Data Bank (PDB) (http://www.rcsb.org/pdb) files that had annotations and treated them as unannotated, using only the protein sequence and coordinates. The idea was to assess how well ProKnow could recover the annotations using jackknife-like criteria (Table 3). Of the two sets, set A had all categories of annotation, while set B excluded electronically evidenced ones (Tables 1 and 2). The separate sets were created to see if electronically evidenced GO terms affect ProKnow performance. These electronic annotations are a majority in the knowledgebase and are less reliable. The quality of assignments estimated by ER varied between 0–6, with 82% of the assignments within the range of 3–5 for set A and 68% in the range of 1–3 for set B. ER values 4–6 indicate major contribution from electronically evidenced GO terms. Neither the ER nor CC parameters showed a clear correlation with ProKnow performance: the fraction of correct assignments was not dependent on either ER or CC values.

The DAG structure of the GO dictionary allows quantitative interpretation of the precision of each assignment of a GO term. To make this quantification, a GO term and all its parent terms need to be drawn as a DAG based on the relationships described by the GO dictionary. We call this DAG of the GO term and its parent terms a PDAG. All GO terms in the PDAG of the assignment and the PDAG of the PDB annotation can then be compared by pairwise matching. For no matching GO terms between the PDAGs, an assignment is marked false positive (FP). If there is a match (called true positive, TP), the location of the matching GO terms can be noted by counting the number of edges to the terms from the root term. Each traceable path from the root term is called a full ontology. Sometimes, however, there may be more than one traceable path from the root term to the required GO term. Here, we select the path with maximum number of edges to root and note it as the ontology depth of the assignment. The ontology depth indicates the descriptive level of the assigned function (example: PDAG::depth = enzyme → hydrolase → ATPase:: n → n + m → n + m + p, where n is the maximum number of edges connecting enzyme from the root term [GO:0003674 for molecular function], and m and p for enzyme to hydrolase and hydrolase to ATPase, respectively). To quantify the rank of performance ranging from total failure (value = 0) to a complete success (value = 1), we defined another parameter called assignment specificity [TP/(TP+FP)].

Table 3. Overview of the Assignments Made by ProKnow Using Jackknife-like Criteria

Set	No. of proteins	No. of PDB GO Molecular Function Annotations	No. of GO Molecular Function Assignments by ProKnow	Percent Assignments with CC $\geq$ 8
A	1507	4455	9598	89%
B	383	527	2509	56%

Set A has all categories of annotation, while set B excluded electronically evidenced ones. The electronic annotations are a majority in the knowledgebase and are less reliable.



The overall ProKnow performance was assessed based on the variation of the assignment specificities at various ontology depths.

The ability of ProKnow to make useful annotations can be judged from variation of assignment specificity with ontology depths (Figure 2A). Eighty-nine percent of PDAG assignments have at least one GO term match with annotated PDAGs. As we go down the ontology depths, the specificity decreases sharply to around 0.6 for depth 2 and to around 0.4 for depth 9. A deep assignment is more difficult, as is evident from the general DAG structure for all ontologies (Supplemental Figure S1B). The repeat analysis with set B shows a similar distribution. The assignment specificity is diminished due to the smaller size of the ProKnow knowledgebase used for querying set B compared to set A. That the assignment specificity of ProKnow is not significantly diminished with increasing ontology depths is evident from the nonexponential nature of the specificity curve in Figure 2A.

A receiver-operator plot allows us to estimate the efficacy of various BWs in filtering out false assignments. In Figure 2B, 12 BW thresholds of 1.0, 0.80, 0.60, 0.40, 0.20, 0.15, 0.10, 0.05, 0.04, 0.02, and 0.01 are used. For each of these thresholds, we count how many TP and FP assignments have been made by ProKnow. The plot of these counts shows that the performance of BWs in ProKnow is very efficient. A perfect receiver-operator plot for any BW would show vertical lines. The slopes

Figure 2. The Statistical Evaluation of Assignment Performance

A true-positive assignment is indicated by TP, and a false postive is indicated by FP. Ontology depth indicates the description level of the assignment made; it is calculated by counting the maximum number of edges connecting the root term (GO:0003674 for molecular function) to the given GO term. The main plots refer to set A, while the insets refer to set B.

(A) The fraction of correct assignments (left y axis) at each ontology depth, also termed the assignment specificity (shown by the black dots). The number of such assignments made at each ontology depth is shown as a bar graph (right y axis). That ProKnow performance is not significantly diminished with increasing ontology depths is evident from the nonexponential nature of the assignment specificity curve.

(B) The receiver-operator plot showing the fraction of TP and FP using Bayesian weight as thresholds. The Bayesian weight thresholds used in the plot are 1.0, 0.80, 0.60, 0.40, 0.20, 0.15, 0.10, 0.05, 0.04, 0.02, and 0.01. At each of these thresholds, the TPs and FPs were counted having Bayesian weight within the threshold value. The data are shown only for set A. The steep slope of the curves indicates that Bayesian weights are very effective in discriminating for correct assignments.

(C) Plot indicating the coverage (i.e., fraction of all annotations for the PDB files in the test set) achieved (left y axis) by ProKnow at various ontology depths. The bars show the number of annotations present at each depth in the test set (right y axis). A maximum of 93% coverage was achieved. The lowest specificity of recovery is 0.6, meaning that a considerable proportion of the annotations were recovered by ProKnow, irrespective of the ontology depth.

(D) The precision of the ProKnow algorithm in recovering exact annotation. Around 70% of the annotations could be recovered exactly as they are present in the database (Δ Ontology Depth = 0). The rest were imprecise by one or more edges in the PDAG, the fraction of these assignments decreasing with the increasing number of edge differences.

of the curves for all ontology depths are very steep, indicating rapid increase in filtering power with increasing BWs. However, the decrease in the slope for lower BWs evident for depths 2–9 suggests considerable decrease in filtering efficiency at lower BWs. This is due to the larger number of assignments that must be screened at lower BWs compared to higher BWs. The larger number of assignments at lower BWs can be rationalized from the average number (~6) of assignments per protein in the test set (Table 3): because the sum of the assigned BWs is restricted to 1, the distribution of the BWs is therefore more often restricted to lower values for proteins with higher numbers of assignments.

The fraction of GO terms in the PDAGs of the original PDB annotations recovered by ProKnow gives an estimate of the coverage achieved (Figure 2C). The plot shows 93% correct coverage for at least one match for the GO terms in the annotated PDAGs. The specificity of assignment is greater than 0.6 for ontology depths less than 7. This is true for both test sets A and B. The high levels of overall coverage indicate that the algorithm is able to recover correctly a large majority of the original PDB annotations.

We also evaluated how many times ProKnow assigned precisely the same GO term to a protein as in the database, and if it did not, by how many edges it erred in the PDAG (Figure 2D). A zero difference in the number of edges means an exact assignment made. The curve shows that approximately 70% of the GO terms have been assigned correctly for proteins in set A, the value being marginally lower for set B. In general, there are fewer annotations for the test set proteins having deep ontologies (Figure 2C), and as a result there is not much scope for the ontology depths of annotated and assigned functions to differ by a large number of edges.

Statistical Significance

We estimate the statistical significance of our results by assuming the null hypothesis: the prediction scheme is better in assigning a GO term from the "protein features" (sequence/fold, etc.) than random selection of function by simply choosing the GO term in proportion to the frequency with which it occurs in the ProKnow knowledgebase. Z scores calculated based on this hypothesis suggest assignments made at ontology depth > 1 are statistically significant (Supplemental Figure S2).

Sequence-Only Assignments

We applied ProKnow to the 3999 gene sequences in the *Mycobacterium tuberculosis* (TB) H37Rv genome. Here, ProKnow used the top 50 fold recognition hits from DASEY (Mallick et al., 2002) for mapping the fold-based annotation profiles from the knowledgebase. RIGOR (Kleywegt, 1999) was turned off in absence of three-dimensional coordinates, lowering the maximum CC value by 2. As the majority of the genes in the TB genome lack functional annotation, the ProKnow assignments could not be evaluated directly. ProKnow assigned at least one functional term to 97% of the genes at various confidence levels (Figure 3). If we look at assignments that are reasonably accurate (BW $\geq$ 0.4 and CC > 4), the coverage is around 50%, which is comparable to HMM and better than BLAST. Currently, an HMM-based search on the TB genome using PFAM-B domains (Bateman et al., 2004) finds hits for around 42% of the genes at a statistical significance value better than e-03. The coverage using BLAST on annotated sequences is significantly lower. We expect the bulk of the ProKnow assignments of molecular function and biological process GO terms at ontology depth 5 or deeper to be of practical use. The results for all the genes in the TB genome, their function-based similarity, and links can be explored at http://www.doe-mbi.ucla.edu/Services/ProKnow/biolatlas.php.

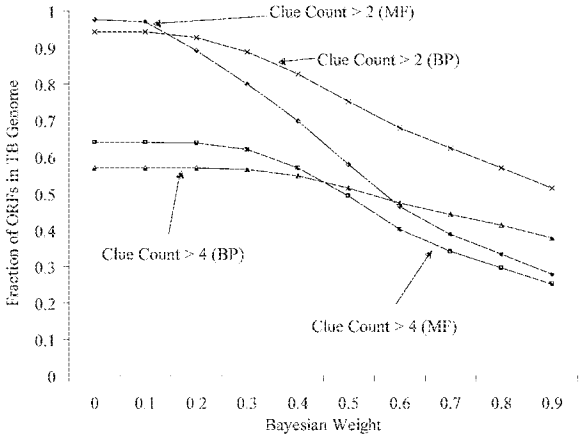


Figure 3. The Distribution of the Bayesian Weights of the Assigned GO Terms for the ORFs in the *Mycobacterium tuberculosis* Genome

The assignments were derived from the knowledgebase containing all categories of annotations, including electronic annotations. At least one GO term is assigned to 97% of the genes in the TB genome. Around 50% of the genes have been assigned GO terms at a high confidence level (Bayesian weight $\geq$ 0.4 and clue count > 4).

Functionally linked proteins are more likely to be part of a single biological process. To check if this is evident from ProKnow biological process assignments, we compared examples of ProKnow-derived biological process assignments with clusters of proteins inferred by combined functional linkage methods (Strong et al., 2003). We found many new groups of proteins having a common biological process not described by linkage methods. For example, Rv2029c, Rv2202c, and Rv2436, involved in ribose metabolism, are assigned to a high confidence (BW = 1, CC $\geq$ 4) (Table 4). BLAST searches and searches against the cluster of orthologous groups (COG) (Tatusov et al., 1997) corroborated their putative involvement in carbohydrate transport and metabolism. Despite the lack of many matches between ProKnow and linkage methods, some biological processes do match well. One such assignment is molybdopterin cofactor biosynthesis (GO:0006777) to 17 assigned genes from TB (Table 4). The genes shown in bold in Table 4 matched linkage method assignments. Of the unmarked genes, three genes (Rv0438c, Rv0866, and Rv3323c) were assigned at high levels of confidence (BW $\geq$ 0.4 and CC $\geq$ 4). Their functions were also substantiated through COG database searches and annotations derived through BLAST. Only two functionally linked genes predicted by linkage methods (Rv3116

Table 4. Two ProKnow-Assigned Representative Examples of Biological Processes and the TB Genes Involved in Them

	Gene Name	Bayesian Weight (BW)	Clue Count (CC)	Homology Annotation	COG Function
Ribose metabolism	Rv0628c	0.004	2	---	
	Rv2029c	1	6	PfkB	
	Rv2202c	1	4	CbhK	COG0524: carbohydrate transport and metabolism
	Rv2436	1	6	RBSK	COG0524: carbohydrate transport and metabolism
	Rv2542	0.03	4	---	
Molybdopterin cofactor biosynthesis	Rv0416	0.030	4	this	COG2104: coenzyme metabolism
	Rv0438c	1	6	moeA2	COG0303: coenzyme metabolism
	Rv0476	0.012	2	---	
	Rv0864	1	7	moaC2	COG0315: coenzyme metabolism
	Rv0865	0.999	4	mog	COG0521: coenzyme metabolism
	Rv0866	1	5	moaE2	COG0314: coenzyme metabolism
	Rv0869c	1	4	moaA2	COG2896: coenzyme metabolism
	Rv0984	1	4	moaB2	COG0521: coenzyme metabolism
	Rv0994	1	6	moeA	COG0303: coenzyme metabolism
	Rv1443c	0.10	2	---	
	Rv1498A	0.004	2	---	COG0028: amino acid transport and metabolism
	Rv3109	1	4	moaA	COG2896: coenzyme metabolism
	Rv3111	1	6	moaC	COG0315: coenzyme metabolism
	Rv3119	0.96	5	moaE	COG0314: coenzyme metabolism
	Rv3323c	0.69	5	moaX	COG1977: coenzyme metabolism
	Rv3324c	1	6	moaC3	COG0315: coenzyme metabolism
	Rv3843c	0.004	2	---	

The first process, "ribose metabolism," is defined as the chemical reactions and physical changes involving D-ribose (ribo-pentose). The second, "molybdopterin cofactor biosynthesis," is defined as the formation from simpler components of molybdopterin cofactor (Moco), essential for the catalytic activity of some enzymes, e.g., sulfite oxidase, xanthine dehydrogenase, and aldehyde oxidase. To assign GO terms for the biological processes to the genes, ProKnow extracted their protein features, which gave clues that were analyzed by Bayes' theorem to output Bayesian weights (BW), indicating probability of occurrence of those functions. A weight from a protein feature is a clue, and the total number of weights from the extracted protein features for evaluating a biological process is designated as clue count (CC). BLAST pairwise sequence comparison to nonredundant sequence database gave the homology annotation. The nonredundant sequence database is a collection of all published protein sequences that do not share more than 95% sequence identity. A similar comparison against the database of orthologous sequences gave the COG function. The genes predicted for molybdopterin cofactor biosynthesis that match with the combined linkage map of TB (Strong et al., 2003) are shown in bold. Notice that the homology annotation and the COG function agree with the ProKnow assignments, especially when BW $\geq$ 0.4 and CC $\geq$ 4 (italicized). Some of these high-confidence predictions are not detected by the linkage methods.

and Rv3206c) are not assigned to molybdopterin cofactor biosynthesis by our method: Rv3116 is assigned GO:0006118 for electron transport and Rv3206c as GO:0006464 for protein modification. A look into the combined PDAG of GO:0006777, GO:0006118, and GO:0006464 showed that these function are not totally unrelated. In the PDAG, GO:0042558: pteridine and derivative metabolism is a common parent GO term linking GO:0006118 to GO:0006777 for Rv3116; GO:0009058: biosynthesis links GO:0006464 to GO:0006777 for Rv3206c. Thus, it is likely that all of these open reading frames (ORFs) may in some way be involved in a common biological process.

Individual Examples of Functional Assignment

We tested ProKnow on protein pairs that are enzyme-nonenzyme homologs (Todd et al., 2002) (Table 5). These proteins share the same fold with varying degrees of sequence identity and have diverged to an extent where, despite an ability to bind a substrate, they lack functional machinery for catalytic reactions. Assignments of ProKnow molecular function GO terms for these proteins were individually evaluated by looking at annotations already present for the PDB file and descriptions compiled by Todd et al. (2002). Most top-ranked predictions from ProKnow are correct, although in some cases the description of function is not to the desired detail. For example, PDB 1dps, which is a DNA protection molecule, is assigned a binding activity (GO:0005488) at 0.65 BW—only broadly correct. Similarly, Cre recombinase (PDB 1crx) is assigned GO:0003677 for DNA binding, but recombinase activity is not obvious from its PDAG. The only assignment completely false is for PDB 1ndo, a noncatalytic naphthalene dioxygenase assigned as an enzyme. It appears that the C-terminal region that blocks the active site is not able to contribute in any way toward a proper assignment. Another interesting aspect of functional divergence is evident from the comparison of PDB 1a73 and 1mhd, which are an endonuclease (GO:0004519) and a DNA binding transcription regulator (GO:0003677), respectively. Evaluation of the PDAG for the GO terms shows that DNA binding is a parent term of endonuclease activity. This suggests that homologous proteins with common residual function may share a part of the ontology tree.

Discussion

The sequence of a protein encodes all information required for its fold and function, but we are not always able to decipher the function from sequence or structure alone. ProKnow assigns function by extracting and

Table 5. ProKnow Molecular Function Assignment for Enzyme-Nonenzyme Homologs

Enzyme		Nonenzyme			
PDB Code	GO Terms (BW)	PDB	GO Terms (BW)	Reason for Loss/Gain of Activity	Functional Similarity
1a73 (endonuclease I-Ppol)	0004519 (1)	1mhd (Smad transcription regulator, MH1 domain)	0003677 (1)	R61deleted; general base mutated H98A Mg^{2+} cofactor binding residue deleted: N119	both bind DNA
1xik (ribonucleotide reductase)	0004748 (0.71)	1dps (DNA protection molecule)	0005488 (0.65) 0008199 (0.35)	enzyme di-iron site absent	none
	0016491 (0.28)	2fha (ferritin)	0005488 (0.5) 0008199 (0.5)	The di-iron site is absent in ferritin light chain but present in heavy chain.	ferroxidase activity of di-iron site
1pda (porphobilinogen deaminase)	0004418 (0.54) 0016829 (0.44)	1ixh (phosphate binding protein)	0003723 (0.96) 0003676 (0.04)	A wide variety of water-slouble ligands, such as mono- and oligosaccharides, amino acids, ologopeptides, and sulphate and phosphate are bound by periplasmic binding domains.	substrate binding
1crx (Cre recombinase)	0003677 (0.996) 0005524 (0.002)	1bl0 (MarA transcriptional activator)	0003677 (0.79) 0003700 (0.21)	Two repeats of the homeodomain-like module containing the DNA binding helix-turn-helix motif are shared.	DNA binding
1qjg (ketosteroid isomerase)	0004769 (0.89)	1oun (nuclear transport factor 2)	0008565 (0.998) 0003723 (0.002)	The proteins do not share only one catalytically essential residue in common.	Both the proteins bind the aromatic groups in equivalent hydrophobic cavities.
	0003824 (0.10)	1ndo (napthalene dioxygenase noncatalytic β subunit)	0003824 (0.97) 0005215 (0.03)	The C-terminus fills the region equivalent to the enzyme active site cavity.	
1aoz (L-ascorbate oxidase)	0005507 (0.80) 0016491 (0.20)	1nwp (azurin)	0005507 (0.63) 0005489 (0.36)	Multicopper oxidases have different types of copper sites that make them catalytically active.	Cu type I site for single electron transfer oxygen and activation of dicopper site
1bug (catechol oxidase)	0016491 (1)	1oxy (hemocyanin)	0005344 (0.57) 0016491 (0.40)	The Phe residue in the N-terminal domain aligns itself to block access of substrates, allowing 1oxy to function as an oxygen transporter.	

The Bayesian weight (BW) for each assigned GO term is given in parentheses. Those GO terms which exactly match with the PDB GO terms in the database are shown in boid. PDB codes 1a73, 1mhd, 1pda, and 1oxy did not have any previous annotation in the database and are new molecular function assignments. Most of the top hits in the table match correctly with the function described in literature for the protein, except for 1ndo. Some of the predictions are not to the desired detail, such as in 1crx and 1dps, where binding is predicted without the activity associated with it. Notice that despite obvious similarities between the enzyme and nonenzyme homologs, the functional predictions are quite accurate. The complete table with all ProKnow assignments can be found in Supplemental Table S2.
Descriptions of GO terms in this table: GO:0003676, nucleic acid binding; GO:0003677, DNA binding; GO:0003700, transcription factor activity; GO:0003723, RNA binding; GO:0003824, enzyme activity; GO:0004418, hydroxymethylbilane synthase activity; GO:0004519, endonuclease activity; GO:0004748, ribonucleoside-diphosphate reductase activity; GO:0004769, steroid delta-isomerase activity; GO:0005215, transporter activity; GO:0005344, oxygen transporter activity; GO:0005488, binding; GO:0005489, electron transporter activity; GO:0005507, copper ion binding; GO:0005524, ATP binding; GO:0008199, ferric iron binding; GO:0008565, protein transporter activity; GO:0016491, oxidoreductase activity; GO:0016829, lyase activity.

interpreting protein features from sequences and structures. Most servers that annotate protein function do so on the basis of homology, which has commonly been interpreted for similarity in function. Of the few "function" annotating servers, Protfun (Jensen et al., 2003) takes sequences alone and predicts for probability among 14 broad functional classes, such as transporter, growth factors, transcription factors, etc. Another sequence-based server, Wilma (Pric et al., 2004), has somewhat similar goals but is implemented using a different algorithm. For both of these servers, as for ours, metaserver strategies have been used, but our approach differs by implementing a knowledgebase of annotation profiles coupled with Bayesian scoring. The combined advantage of using the GO term profiles for protein features and Bayes' theorem extends the coverage on assigning function beyond what is currently available.

The capability of ProKnow is highlighted by its efficient annotation performance and ability to distinguish enzyme-nonenzyme pairs despite obvious similarities in sequence and structure between the homologous

A

DALI

PDB+chain	GO terms	Z-score
1CSXA	F_1, F_3, F_4, F_8	8.5
1TUV_	F_1, F_3, F_5, F_6, F_8	6.3
1SLRA	F_1, F_2, F_3, F_6, F_8	3.4
1AUV_	F_1, F_3, F_8, F_{10}	2.4

(for details Holm & Sander, 1998)

B

RIGOR

3D-motif	GO terms	M-score
9AGA3	$F_1, F_2, F_4, F_8, F_{10}$	0.5
3ASY4	F_1, F_2, F_3, F_6, F_8	0.3
1CSX2	F_1, F_3, F_4, F_8	0.4
1GSX1	F_1, F_3, F_8, F_{12}	0.1

M-score = $(\sum C_i)/$ (n * Max(C_i) * rmsd); rmsd = average root mean square deviation, minimum rmsd is set to 0.01 Å; n=number of pseudoatoms in the motif; C_i as in box D; summation is over n atoms in motif (see details Kleywegt, 1999).

C

PSI-BLAST

Seq. ID	GO Terms	M-eval
0012131	$F_1, F_2, F_3, F_8, F_{12}$	0.995
P567456	F_1, F_3, F_8, F_{10}	0.895
0012345	F_1, F_3, F_5, F_6, F_8	0.685

M-eval = 1 – e-value; (for details Altschul, 1997)

D

PROSITE

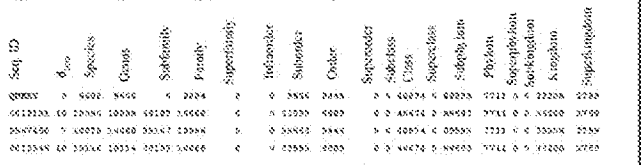
Motif	GO Terms	P-score
GG. ..GG	$F_1, F_3, F_6, F_8, F_{18}$	1.675
AG. ..AA	F_1, F_3, F_4, F_8	3.215
DNAAEA	F_1, F_2, F_3, F_7, F_8	0.367

P-score = $\sum C_i$, where C is conservation score of each residue i in the query sequence as derived from PSI-BLAST alignment; summation is over all residues in the motif.

$C_i = [(w_i \ln(w_i))/(N * d_{avg\_taxa})] - \text{Min}(C_i)$

w_i = fraction of identical residues in alignment of N sequences

d_{avg_taxa} = Average taxonomic distance between the sequences having identical residues. Taxonomic distance (d_{taxa}) is defined as the number of hierarchy steps from a shared taxonomy lineage (see below). For a taxonomy level if there is no information available, it is represented by 0 and ignored for comparison.



0012131 . . ARLKRNRQV . .
P567456 . . ARLKRNRQV . .
0012345 . . VPVTRNRQV . .

N=3
w for residue N = 2/3 = 0.667
d_{avg_taxa} = (10 + 7) / 2 = 8.5

E

DIP via PSI-BLAST

Seq. ID	GO terms	M-eval
No hits		

M-eval = 1 – e-value (for details Altschul, 1997 and Xenarios et al., 2002)

Decision Table

F	A		B		C		D		E	Clue Count	Bayesian weight*
	W_1[§]	W_2[¶]	W_3[§]	W_4[¶]	W_5[§]	W_6[¶]	W_7[§]	W_8[¶]	W_9[¶]		
F_1	8.5	4	0.352	4	0.995	3	1.406	3		8	0.422
F_2	3.4	1	0.4	2	0.995	1	0.367	1		8	0.0007
F_3	8.5	3	0.225	3	0.995	3	1.406	3		8	0.152
F_4	8.5	1	0.45	2			3.215	1		6	-
F_5	6.3	1			0.685	1				4	-
F_6	6.3	2	0.3	1	0.685	1	1.675	1		8	0.003
F_7		3					0.367	1		4	-
F_8	8.5	4	0.352	4	0.995	3	1.406	3		8	0.422
F_{10}	2.4	1	0.5	1	0.895	1				6	-
F_{12}			0.1	1	0.995	1				4	-
F_{18}							1.675	1		2	-

[§]Cumulative Average.CA_0=CA_8; CA_1=(CA_0+CA_1)/2; CA_i = ((CA_{i-1}+CA_{i-2})/2) + CA_i)/2; where i is the clue derived from protein feature for function F. [¶]Frequency of occurrence of the GO term in the hit list. [§]Max(score for GO term). For DIP (W_9) only frequencies are used because the GO terms are ported based of the edges and not the node. * using equation 1.

Figure 4. A Sample Evaluation of Bayesian Weight

The boxes labeled A to E correspond to the feature extractors described in Figure 1. Each feature extracted is mapped to GO terms using the annotation profiles from the ProKnow knowledgebase (examples are given in Supplemental Table S1). Z score in box A, M score in B, M-eval in C, P score in D, and M-eval in E are referred to as clues to a function in Equation 1 (see main text). Brief descriptions of how clues are computed are given in the individual boxes. The decision table is a compilation of all the clues and the associated functions with the purpose of choosing the cases with the highest clue count (CC) for weighting by Bayes' theorem using Equation 1 and output as final results.

partners. A major factor contributing to the accuracy of performance of ProKnow is the explicit use of protein domains (Guo et al., 2003) for functional assessments when we have the structural information in hand. In the absence of structural information, ProKnow can make sequence-only assignments. Then, the use of GO vocabulary allows us to bypass the need for domain partitioning (explained in Supplemental Figure S3). This makes ProKnow a useful function annotation tool for ORFs with no domain information. Additionally, the use of fold recognition in the method increases the accuracy of functional assignments.

An important aspect of interpretation of any ProKnow assignment is an understanding of the weights on

which it was inferred (Supplemental Figure S4). A high confidence assignment is one that has BW $\geq$ 0.4, CC > 4, and ER < 5, the order of their importance being BW > CC > ER. Because we are dealing with novel proteins, the BW and CC values may not always be high and therefore not of best confidence. In every case, we allow the user to check all the protein features from which the annotation was derived by ProKnow. For example, in screening enzyme-nonenzyme proteins, one would expect DALI to be less effective in discriminating functions, and therefore a look into the protein features helps to know whether PSI-BLAST, PROSITE, RIGOR, or DIP is the basis for discrimination. Sometimes, however, ProKnow may fail to detect any signal for a function from the knowledgebase because of the extreme novelty of the protein. In that case, ProKnow outputs a large number of GO terms, most of which are noise. In such cases, the user can use the relationships from the GO dictionary to merge functions by manually locating assignments that share a common parent node in the PDAG. This can reduce the large pool of GO terms to a smaller number, allowing for a better and more confident assessment of function. In practice, we expect molecular functions to be predicted more confidently than biological processes, because features of a protein are more intimately linked to its function at the biochemical level rather than the larger biological function to which it contributes.

Concluding Remarks

We have developed ProKnow for annotating protein structure using the controlled vocabulary of the Gene Ontology dictionary. The method integrates various programs, such as PSI-BLAST, PROSITE, DALI, and RIGOR, to extract similarity of the query protein to protein features in the ProKnow knowledgebase. These features include sequence, fold, motifs, and functional linkages. The annotation profile of features stored in the precompiled knowledgebase is used to map features to functions. The likelihood of the function is derived using Bayesian scoring by updating weights obtained from individual protein features. In this scheme, functions linked to a maximum number of protein features are used for scoring. The final output is a list of functions and their Bayesian weights. The evaluation of our method gave a specificity of ~0.89 at ontology depth 1 and 0.4 at depth 9; the coverage was 93%. Around 70% of the annotations were assigned correctly. The architecture of our method also allows us to predict function from sequence alone. An application of ProKnow to the TB genome shows that ProKnow is able to assign around 50% of genes in the genome with high confidence. We also tested the method on enzyme-nonenzyme homologous partners with distinct function, where the method detected the majority of functional dissimilarities. Our prediction server is available for use, and we hope it will assist the scientific community in their quest to understand protein function.

Experimental Procedures

We assume that a protein has a set of functions F_1, F_2,...F_n, for which there exists evidence given by Bayesian weights BW_1, BW_2...BW_n. BW is based on the clues extracted from sequence, fold, active site geometry, etc., which we call "features" of the protein. An individual clue from a "protein feature" is used to relate the extracted protein feature to the features in the ProKnow knowledgebase to get the likelihood of the functions. The total number of extractable clues from protein features for a function F_n is designated as clue count (CC; maximum value is 9 for a structure query). The higher the CC, the more confident we are in the BW for the function (this assumption breaks down when the clues are not mutually exclusive). During query, ProKnow assigns numerous annotation profiles to the protein from the ProKnow knowledgebase based on features; we choose only those functions from the assigned profiles that have the maximum CC. The likelihoods of these functions are analyzed by Bayes' theorem (Pitman, 1997) to arrive at the best-evidenced set of functions:

$$p(F_n \mid \text{clue}) = p(F_n) \times p(\text{clue} \mid F_n) / Z. \quad (1)$$

The left-hand side of the equation $p(F_n|\text{clue})$ is called the Bayesian posterior probability given a clue from a protein feature for function F_n. The right-hand side numerator is the product of the prior probability of the protein having the function, $p(F_n)$, and the probability of the clue given a function, $(\text{clue}|F_n)$. The denominator Z is a normalization factor [$Z = \Sigma\, p(F) \times p(\text{clue}|F)$]; essentially a summation of the numerator over all functions, F_n.

Every time a probability of a clue given a function [$p(\text{clue}|F_n)$] is input into Equation 1, the formula returns a posterior probability for the occurrence of that function based on the associated prior probability (equiprobable in the first step). The posterior probability is used as input as prior probability for the next evaluation of likelihoods, and the equation is repeatedly used until all clues are analyzed. This gives a set of functions and the final BWs. A sample ProKnow assignment is given in Figure 4.

ProKnow Knowledgebase

We have used the SWISS-PROT.GOA file from GO website (http://www.geneontology.org) as our master file. The file contains GO terms associated with each protein sequence. We scanned this file for protein features and associated each feature to the GO terms for the sequence from which the protein feature was extracted. This way, each protein feature was associated with all potential functions it possibly implicates. The ProKnow knowledgebase containing the annotation profile for each protein feature was generated from this single master file (Table 1). The GO dictionary was downloaded separately to generate and analyze the PDAGs; it is not part of the ProKnow knowledgebase. All downloaded files, including the GO dictionary, correspond to the version existing as of June 2003.

The Test Set

To test our method, we chose proteins from the FSSP library (database of proteins with distinct fold derived by the DALI server, http://www.ebi.ac.uk/dali) that had GO entries in our database. A strict jackknife test required that during evaluation we exclude not only self-entries, but all other proteins which are highly similar. For this, we scanned the PDBAA database (which lists all protein chains in the PDB at 95% sequence identity) and noted all similar sequences. These proteins were excluded from the database during jackknife-like evaluations. Because our method does not use sequence similarity alone to assign function, a cut-off at 95% sequence identity seemed adequate for a stringent jackknife criterion. Additionally, we call our test jackknife "like" because we do not compute any weight matrices from our training set; as a result, we do not need to have exclusive training and testing sets. Instead, we eliminate individual sets of required proteins in each round of evaluation.

Supplemental Data

Supplemental Data are available at http://www.structure.org/cgi/content/full/13/1/121/DC1/.

Acknowledgments

We thank the Department of Energy—OBER, Howard Hughes Medical Institute, and National Institutes of Health for support and Rob Grothe for discussions.

Received: August 5, 2004
Revised: October 18, 2004
Accepted: October 20, 2004
Published: January 11, 2005

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., et al. (2003). PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. *31*, 400–402.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

Bork, P., and Gibson, T.J. (1996). Applying motif and profile searches. Methods Enzymol. *266*, 162–184.

Bowie, J.U., Luethy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science *253*, 164–170.

Cai, C.Z., Wang, W.L., Sun, L.A., and Chen, Y.Z. (2003). Protein function classification via support vector machine approach. Math. Biosci. *185*, 111–122.

Devos, D., and Valencia, A. (2000). Practical limits of function prediction. Proteins *41*, 98–107.

Eddy, S.R., Mitchison, G., and Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. J. Comput. Biol. *2*, 9–23.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000). Protein function in the post-genomic era. Nature *405*, 823–826.

Gene Ontology Consortium(2001). Creating the gene ontology resource: design and implementation. Genome Res. *11*, 1425–1433.

Gerlt, J.A., and Babbitt, P.C. (2000). Can sequence determine function? Genome Biol. *1*, 1–10.

Gribskov, M., McLachlan, M., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA *84*, 4355–4358.

Guo, J.T., Xu, D., Kim, D., and Xu, Y. (2003). Improving the performance of DomainParser for structural domain partition using neural network. Nucleic Acids Res. *31*, 944–952.

Henikoff, J.G., Greene, E.A., Pietrokovski, S., and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. *28*, 228–230.

Holm, L., and Sander, C. (1998). Touring the fold space with DALI/FSSP. Nucleic Acids Res. *26*, 316–319.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., and Bairoch, A. (2004). Recent improvements to the PROSITE database. Nucleic Acids Res. *32*, D134–D137.

Jeffery, C.J. (1999). Moonlighting proteins. Trends Biochem. Sci. *24*, 8–11.

Jensen, L.J., Gupta, R., Staerfeldt, H.-H., and Brunak, S. (2003). Prediction of human protein function according to Gene Ontology Categories. Bioinformatics *19*, 635–642.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). A new approach to protein fold recognition. Nature *358*, 86–89.

Karp, P.D. (1998). What do we know about sequence analysis and sequence databases. Bioinformatics *14*, 753–754.

Kleywegt, G.J. (1999). Recognition of spatial motifs in protein structures. J. Mol. Biol. *285*, 1887–1897.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol. *307*, 1487–1502.

Mallick, P., Weiss, R., and Eisenberg, D. (2002). The directional atomic solvation energy: An atom-based potential for the assignment of protein sequences to known folds. Proc. Natl. Acad. Sci. USA *99*, 16041–16046.

Pitman, J. (1997). Probability (New York: Springer).

Pric, A., Dominigues, F.S., Lackner, P., and Sippl, M.J. (2004). Wilma-automated annotation of protein sequences. Bioinformatics *20*, 127–128.

Rigoutsos, I., Huynh, T., Floratos, A., Parida, L., and Platt, D. (2002). Dictionary-driven protein annotation. Nucleic Acids Res. *30*, 3901–3916.

Rost, B. (2002). Enzyme function less conserved than anticipated. J. Mol. Biol. *318*, 595–608.

Rost, B., and Valencia, A. (1996). Pitfalls of protein sequence analysis. Curr. Opin. Biotechnol. *7*, 457–461.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Ofran, Y. (2003). Automatic prediction of protein function. Cell. Mol. Life Sci. *60*, 2637–2650.

Schmitt, S., Kuhn, D., and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. J. Mol. Biol. *323*, 387–406.

Strong, M., Graeber, T.G., Beeby, M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (2003). Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. Nucleic Acids Res. *31*, 7099–7109.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). The genomics perspective on protein families. Science *278*, 631–637.

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. *333*, 863–882.

Todd, A.E., Orengo, C.A., and Thornton, J.M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. Structure *10*, 1435–1451.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci. *5*, 1001–1013.

Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N., and Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J. Mol. Biol. *326*, 955–978.

Whisstock, J.C., and Lesk, A.M. (2003). Prediction of protein function from protein sequence and structure. Q. Rev. Biophys. *36*, 307–340.

Wise, E., Yew, W.S., Babbitt, P.C., Gerlt, J.A., and Rayment, I. (2002). Homologous $(\beta/\alpha)8$-barrel enzymes that catalyze unrelated reactions: orotidine 5′-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase. Biochemistry *41*, 3861–3869.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., and Eisenberg, D. (2002). DIP, database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. *30*, 303–305.

Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L., and Lichtarge, O. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. J. Mol. Biol. *326*, 255–261.